ALT-J/M
A prototype Japanese-to-Malay Translation System

Kentaro Ogura, Francis Bond and Yoshifumi Ooyama
NTT Communication Science Laboratories, Kyoto, 619-0237, JAPAN
{ogura,bond,ooyama}@cslab.kecl.ntt.co.jp

Abstract

In this report we introduce **ALT-J/M** — a prototype Japanese-to-Malay translation system. The system is a semantic transfer based system that uses the same translation engine as **ALT-J/E,** a Japanese-to-English system.

1 Introduction

In this report we introduce **ALT-J/M** — a prototype Japanese-to-Malay translation system.

In Malaysia, there has been a lot of work on Malay-English machine translation systems, culminating in Zaharin et al.'s (1994) user-driven machine translation system SISTEP/JEMAH. This is a fully functional translators workbench, with the ability to perform on-line dictionary look-up, glossing and automatic translation of phrases and clauses. It is, however, only available for Malay and English.

In Japan, there has been much work on Japanese-English machine translation, but, none on Japanese-Malay so far. At the NTT Communication Science laboratories, we have been developing **ALT-J/E,** the Automatic Language Translator — Japanese to English since 1987 (Ikehara et al. 1987; Ikehara et al. 1991). The aim is to produce a high quality machine translation system that can be used to facilitate communication via machine translation.

We began work on a Japanese-to-Malay system for two reasons: (1) Malay is from a different language family to both Japanese and English, so it is an interesting test of our translation engine; (2) There is increasing trade between Japan and Malaysia so the need for such a system has been increasing. Both Malay and Japanese are the official language of over 100 million speakers (Crystal 1987:287).

This report is organized as follows. In the next section, we describe the **ALT-J/M** system's architecture. We then look at the lexicon, the backbone of any MT system, in Section 3. In Section 4 we look at some examples of how the system translates. We then look at future issues in Section 5.

2 Architecture

We use the multi-level translation method in our MT system (ALT-J/M), as shown in Figure 1. The input is separated into tense and modal information (the 'subjective' part) and the kernel sentence (the 'objective' part). The objective part is translated using the multi-level transfer method: wide ranging rules, such as direct parse tree transfer are applied first, followed by idiomatic expressions then patterns from the semantic valency dictionary, and finally the default general patterns. This allows transfers of varying granularity.

The process of translation can be divided into seven parts. First, **ALT-J/M** splits the Japanese text into morphemes. Second, it analyses the sentence syntactically, often giving multiple possible interpretations. Next, it rewrites complicated Japanese expressions into more easily translated ones. Fourth, ALT-J/M semantically evaluates the various interpretations. Fifth, syntactic and semantic criteria are used to select the best interpretation. Sixth, this interpretation is used as input to generate Malay. Finally, the Malay sentence is adjusted to give the correct inflectional forms.

Both Japanese and Malay have quite free word orders, but the default orders are very different. Japanese sentences are typically SOV, with the subject followed by the object and the verb final, whereas Malay is SVO, like English. Within the noun phrase, in Japanese, modifiers come before the head noun (like in English), whereas in Malay they typically follow the head.

ALT-J/M runs under Unix and is written mainly in LISP and C. Currently the translation rate is between 5,000 and 10,000 words an hour. As it is still an experimental system it has not been optimized for speed.

Our original system **ALT-J/E,** was not designed to quickly ramp-up to new languages like, for example, the BOAS system (Nirenberg & Raskin 1998). However, we were able to produce a functional Japanese-to-Malay prototype in approximately 7 person-months.

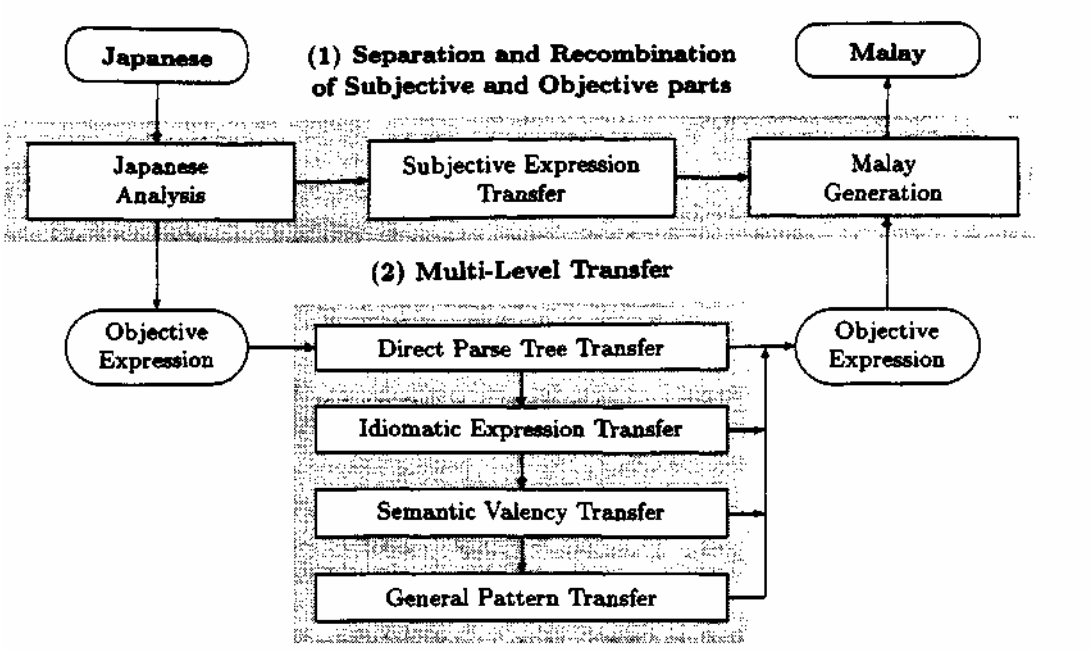

Figure 1: The Multi-Level Translation Method

3 Lexicon

Our dictionaries provide detailed information about the meanings and use of words. Most words can express several concepts. Which concept a word expresses in a text is determined by its relationship with the other words in the text. For example the noun *hotel* is used as a **place** in *I walked straight to the hotel* and as an **organization** in *The hotel was established in 1900.*

Our ontology classifies concepts to use in expressing relationships between words. It represents a kind of common knowledge. Relationships between concepts and words are described in our word semantic dictionaries.

In **ALT-J/M,** the ontology has several hierarchies of concepts: with both **is-a** relationships and **has-a** relationships. 2,800 attributes (12-level tree structure) for common nouns, 200 attributes (9-level tree structure) for proper nouns and 108 attributes for predicates.

3.1 Semantic Word Dictionary

To analyze Japanese sentences a Japanese word semantic dictionary is provided. The dictionary includes 100,000 common nouns, 200,000 proper nouns, 70,000 technical terms and, 30,000 other words: 400,000 words in all (Ikehara et al. 1997).

Figure 2 shows an example of one record of the Japanese semantic word dictionary.

Each record has an index form, pronunciation, a canonical form, syntactic information and semantic attributes. The syntactic information includes the part of speech, inflections, detailed parts of speech, conjunctive conditions and so on. Each word can have up to five common noun attributes and ten proper noun attributes. In the case of *hoteru* "hotel", there are two common noun attributes and no proper noun attributes.

To transfer Japanese words to Malay, a Japanese-to-Malay transfer word dictionary is provided. A simplified example of the entry for *hana* “nose” is given in Figure 3.

Each record of the dictionary has a Japanese index form, a sense number, a Malay index form, Malay syntactic information, Malay semantic information, domain information and so on. Malay syntactic information includes the part of speech, whether a possessive pronoun is required, and so on. This dictionary is also used when the system generates Malay sentences. In addition, the dictionary may have information on selectional restrictions to help choose the correct translation. For example, *hana* “nose” is translated as *hidung* by default, but if it is an elephant's nose then it will be translated as *belalai.*

Our prototype dictionary is still small at present with only about 20,000 Malay entries. However, *if* a Malay translation cannot be found, we use the translation from our English dictionaries. This is better than Japanese, for two reasons: (1) there are many English loan words in Malay where the English is similar to or the same as the Malay e.g. *hotel* “hotel”; (2) most Malay speakers understand more English than they understand Japanese. Our Japanese-English dictionary contains some 400,000 words.

INDEX FORM	ホテル (*hoteru*)
PRONUNCIATION	/hoteru/
CANONICAL FORM	ホテル
PART OF SPEECH	noun
SEMANTIC ATTRIBUTES	COMMON NOUN lodging (⊂ place) enterprise (⊂ organization)

Figure 2: Japanese Lexical Entry

INDEX FORM	鼻 (*hana*)	
SENSE 1	MALAY TRANSLATION	*hidung*
	ENGLISH GLOSS	*nose*
	PART OF SPEECH	noun
	POSSESSED BY DEFAULT	yes
	SEMANTIC ATTRIBUTES	COMMON NOUN nose (⊂ body part)
SENSE 2	MALAY TRANSLATION	*belalai*
	ENGLISH GLOSS	*trunk*
	PART OF SPEECH	noun
	POSSESSED BY DEFAULT	yes
	SEMANTIC ATTRIBUTES	COMMON NOUN nose (⊂ body part)
	SELECTIONAL RESTRICTION	modified-by *zō* "elephant"

Figure 3: Japanese-Malay Lexical Entry

The Malay dictionary is used for the generation of Malay sentences. Each record includes the index form, its part of speech, inflections, and derivational forms. For example, Malay verbs are listed with their stem as index, and the *me-* form as an inflection: e.g. *ambil* "take" ⇒ *megambil.*

3.2 Semantic Structure Dictionary

The basic structure of a sentence comes from the relationship between the main verb and nouns. A structure transfer dictionary is provided for producing basic sentence structures.

We have only 100 or so Malay verbs in our prototype structure dictionary at the moment. As for the word dictionary, if a Malay word cannot be found, we use English. **ALT-J/E** provides 10,000 patterns for the common structure transfer dictionary and 5,000 patterns for the idiomatic structure transfer dictionary. In the common structure transfer dictionary, there is an average of 2.3 patterns for each verb.

Figure 4 shows an example of the common structure transfer dictionary. Each predicate has one or more case-slots associated with it. Each case-slot has information such as grammatical function, case-marker, case-role, semantic restraints on the filler and default order (we don't show all of the features in the examples). We will explain the use of semantic restraints on the filler in lexical selection in the next section.

4 Translation Examples

In this section we give examples of the automatic disambiguation of verbs and nouns. Examples are shown with the Japanese input sentence (in Japanese script) the pronunciation and a gloss, followed by the system's Malay translation and a gloss, then finally an English translation.

4.1 Correct translation of verbs

(1) 私は 休暇を 取る許可を 取り、
watashi-wa kyūka-o toru kyoka-o tori,
I vacation take permission take,
妻は ホテルを 取った。
tuma-wa hoteru-o totta.
wife hotel took.

Saya telah mendapat keizinan untuk mengambil
I past get permission for take
cuti dan isteri saya telah menempah hotel.
vacation and wife I past reserve hotel.

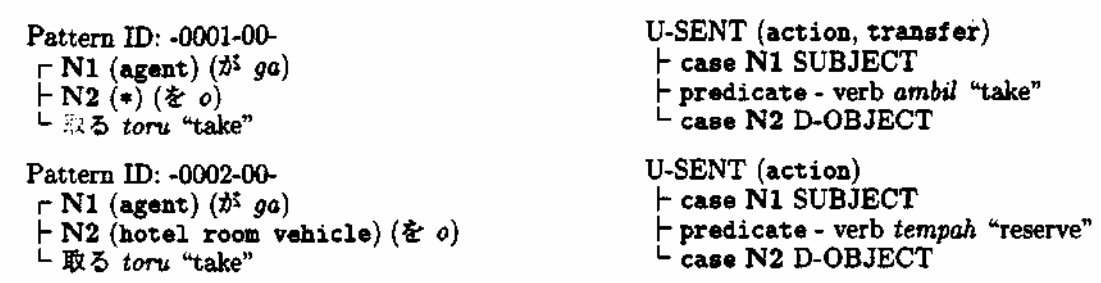

Figure 4: Part of the common structure transfer dictionary for *toru* "take"

'I got permission to take a vacation and my wife reserved a hotel.'

The verb *toru* "take" is used three times in the Japanese sentence. In each case it is translated differently depending on what is taken. **ALT-J/M** has many different patterns it can use to translate *toru.* The three that appear in the example use semantic categories to choose the translation as follows: "take **permission** or **agreement** ..." ⇒ *dapat* "get"; "take **hotel** or **room** or **vehicle** ..." ⇒ *tempah* "reserve"; the default translation is simply *ambil* "take".

Note that *tsuma* "wife" is translated as *isteri saya* "my wife", in the same way as it must be translated as *my wife* in English. In Malay, things that are closely related to people, such as parts of the body and relatives, are normally modified by possessive pronouns. This is not the case in Japanese, where the listener is expected to deduce these relationships from the context. The default assumption, which **ALT-J/M** uses, is that the subject of the sentence is the antecedent of the possessive pronoun (Bond et al. 1995).

4.2 Correct translation of nouns

(2) 象は 鼻が 長い が、豚は 鼻が
zou-wa hana-ga nagai ga, buta-wa hana-ga
elephant nose long but, pig nose
短い。
mijikai.
short.

Gajah panjang belalainya tetapi babi pendek
elephant long trunk.its but pig short
hidungnya.
nose.its.

'Elephants have long trunks but pigs have short snouts.'

In Japanese *hana* "nose" is used for both elephants and pigs, however in Malay it is better to use *belalai* "trunk" for an elephant's trunk. **ALT-J/M**'s Japanese-to-Malay transfer dictionary specifies that *elephant's nose* should be translated as *belalai* "trunk".

This sentence also shows how **ALT-J/M** is able to generate appropriate Malay sentence structures, even when they are different from the Japanese. Japanese allows the 'double subject' construction, when the first subject possesses the second, literally translated as *As for elephants, noses are long* (Oku 1996). In Malay, the possessed entity follows the verb, but must be marked with a possessive pronoun. In this case the third person possessive pronoun *dia* joins onto the proceeding noun as the suffix *-nya.*

5 Discussion and Conclusion

Our prototype Japanese-to-Malay system shows that the framework we developed for Japanese-to-English machine translation can be applied to other target languages. Adding a new language for generation was relatively simple because of our well developed ontology. The word selection rules are all lexically based, so the bulk of the work was in producing the lexicon.

The implementation of our prototype took less than 6 months (7 person-months). Some changes had to be made in the generation system. The rules used to generate English articles *(a/the)* had to be suppressed and we had to make some changes in the default word ordering, in particular in making the noun modifiers follow the noun. These changes were easy to implement in our flexible generation framework based on semantic data structures.

Although Malay does not distinguish between singular and plural, the processing originally developed to generate them in English (Bond & Ogura 1998) was used to generate possessive pronouns, as we showed in our examples. We are planning to introduce a rule to reduplicate the head of noun phrases with generic reference: *budak* "child" ⇒ *budak-budak* "children".

Malay, like Japanese, has a rich pronoun system, with different pronouns used for social superiors or inferiors. Unlike Japanese, it also distinguishes between inclusive and exclusive first person plural pronouns: *kita* "we (including you)" and *kami* "we (excluding you)". We need to build a representation for these uses.

Our next challenge is to increase the size of the lexicons, and test the system with more constructions, such as imperative, interrogative and so forth. We

have been using the Malay-English machine readable lexicon developed by CICC (1994) as a reference. We are currently working on Unking the Malay entries to our Japanese lexicon using the English words as clues. This is similar to the approach of Lafourcade (1997) who combined French and Malaysian via English. We would like to increase our lexicon to several tens of thousands of words, so it can be a useful resource not just for machine translation but for humans or multi-lingual information retrieval (e.g. Hayashi et al.'s (1997) TITAN). In fact, most of the J/E proper noun entries are for Japanese place and person names, and can be used as they are, so we already have over 150,000 entries in our J/M lexicon!

Finally, our Japanese-to-English machine translation system and the Japanese-to-Malay system are separate programs at the moment. We intend to integrate the two systems into a single Japanese-to-X multi-lingual machine translation system which can translate Japanese to English, Malay, Chinese and so on.

Acknowledgment

The authors wish to thank Chengming Guo, Clement Lee, Isamu Shouho, Takefumi Yamazaki and the members of the MT research group for their valuable discussions. We also wish to thank the members of NTT Software Corporation who helped to implement ALT-J/M.

References

Bond, Francis & Kentaro Ogura: 1998, 'Reference in Japanese-to-English machine translation', *Machine Translation,* 13(2-3): 107-134.

Bond, Francis, Kentaro Ogura & Satoru Ikehara: 1995, 'Possessive pronouns as determiners in Japanese-to-English machine translation', in *2nd Pacific Association for Computational Linguistics Conference: PACLING-95,* Brisbane, pp. 32-38, (cmp-lg/9601006).

CICC: 1994, 'Research on Malaysian dictionary', Tech. Rep. 6—CICC—MT54, Center of the International Cooperation for Computerization, Tokyo.

Crystal, David: 1987, *The Cambridge Encyclopedia of Language,* Cambridge University Press.

Hayashi, Yoshihiko, Gen'ichiro Kikui & Seiji Susaki: 1997, 'TITAN: A cross-linguistic search engine for the WWW', in *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval.*

Ikehara, Satoru, Masahiro Miyazaki, Satoshi Shirai & Y. Hayashi: 1987, 'Speaker's recognition and multi-level-translating method based on it', *Transactions of the Information Processing Society of Japan,* 28(12): 1269-1279, (in Japanese).

Ikehara, Satoru, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama & Yoshihiko Hayashi: 1997, *Goi-Taikei — A Japanese Lexicon,* Tokyo: Iwanami Shoten, 5 volumes.

Ikehara, Satoru, Satoshi Shirai, Akio Yokoo & Hiromi Nakaiwa: 1991, 'Toward an MT system without pre-editing − effects of new methods in ALT-J/E −', in *Third Machine Translation Summit: MT Summit III,* Washington DC, pp. 101-106, (cmp-lg/9510008).

Lafourcade, Mathieu: 1997, 'Multilingual dictionary construction and services', in *3rd Pacific Association for Computational Linguistics Conference: PACLING-97,* Meisei University, Tokyo, Japan, pp. 173-181.

Nirenberg, Sergei & Victor Raskin: 1998, 'Universal grammar and lexis for quick ramp-up of MT systems', in *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics: COLING/ACL-98,* Montreal, pp. 975-979.

Oku, Masahiro: 1996, 'Analyzing Japanese double-subject construction having an adjective predicate', in *16th International Conference on Computational Linguistics: COLING-96,* Copenhagen, pp. 865-870.

Zaharin, Yusoff, Tang Engya Kong & See Yee Ling: 1994, 'User-driven machine translation system', in *International Conference on Linguistic Applications,* Universiti Sains Malaysia, Penang, Malaysia, pp. 38-50.